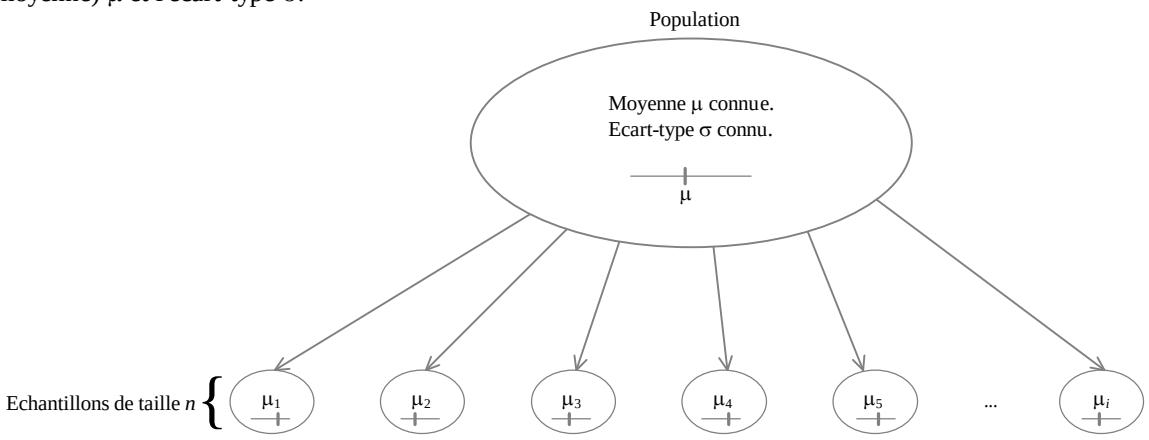
- Partie A - Échantillonnage -

L'objectif de cette partie est de répondre à la problématique suivante : *comment, à partir d'informations (couple moyenne-écart-type ou proportion) connues sur une population, peut-on prévoir celles d'un échantillon ?*

Nous distinguerons deux cas : celui où l'on étudie une **moyenne** dans un échantillon et celui où l'on étudie une **proportion** dans un échantillon.

A.1. Étude de la moyenne d'un échantillon

On dispose d'une population sur laquelle est définie une variable aléatoire X dont on connaît l'espérance (ou la moyenne) μ et l'écart-type σ.



On s'intéresse aux échantillons de taille n. Auront-ils tous la même moyenne ? Non, certains peuvent être constitués d'éléments atypiques et avoir une moyenne très différente de celle de la population (surtout si l'échantillon est de petite taille).

Notons $\overline{X}$ la variable aléatoire qui, à chaque échantillon de taille n, associe sa moyenne ($\overline{X}$ s'appelle encore la *distribution des moyennes des échantillons*). Que peut-on dire de cette variable aléatoire $\overline{X}$?

Théorème Central Limite - Version 1 - (Version faible)

Contexte : variable aléatoire X qui suit une **loi normale** sur la population

$$X \rightsquigarrow N(\mu\ ;\ \sigma)$$

On prélève, au hasard, un échantillon (tirages avec remise[1] ou assimilés) de taille n de moyenne $\overline{X}$.

Alors la variable aléatoire $\overline{X}$ suit également une loi normale :

$$\overline{X} \rightsquigarrow N\left(\mu\ ;\ \frac{\sigma}{\sqrt{n}}\right)$$

Atténuation de la dispersion par le processus d'échantillonnage.

[1] Un tirage avec remise est encore appelé "tirage non exhaustif". Si on fait un tirage sans remise (tirage exhaustif), on modifie la taille de la population au fur et à mesure des tirages, ce qui compliquerait les calculs (intervention d'un facteur d'exhaustivité). Ceci dit, pour des grandes populations le tirage sans remise s'assimile à un tirage avec remise.

Démonstration :

Notons $E = \{x_1 ; x_2 ; ... ; x_n\}$ un échantillon de n éléments prélevés au hasard dans la population.

Pour tout i compris entre 1 et n, notons X_i la variable aléatoire correspondant à la valeur du i-ème élément x_i de l'échantillon. Nous savons, par hypothèse, que :

$$E(X_i) = \mu \text{ et } \sigma(X_i) = \sigma$$

La moyenne $\overline{X}$ des n valeurs de l'échantillon est :

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

D'après les propriétés de la loi normale, nous savons qu'une combinaison linéraire de variables aléatoire qui suivent la loi normale est encore une variable aléatoire qui suit la loi normale. Comme chaque variable aléatoire X_i suit ici la loi normale $N(\mu, \sigma)$, la variable aléatoire moyenne $\overline{X}$ suit donc également une loi normale. Calculons ses paramètres.

D'après la propriété de linéarité de l'espérance :

$$E\left(\overline{X}\right) = \frac{E(X_1) + E(X_2) + ... + E(X_n)}{n} = \frac{n\mu}{n} = \mu$$

D'après les propriétés de la variance :

$$V\left(\overline{X}\right) = \frac{V(X_1) + V(X_2) + ... + V(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

D'où : $$\sigma\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}}$$

Théorème Central Limite - Version 2 - (Version forte)

Contexte : variable aléatoire X qui suit une **loi quelconque** sur la population avec $E(X) = \mu$ et $\sigma(X) = \sigma$.

On prélève, au hasard, un échantillon (tirages avec remise ou assimilés) de taille n, **avec $n \geqslant 30$**, de moyenne $\overline{X}$.

Alors la variable aléatoire $\overline{X}$ suit **approximativement** une loi normale :

$$\overline{X} \rightsquigarrow N\left(\mu ; \frac{\sigma}{\sqrt{n}}\right)$$

Ce théorème dû aux mathématiciens *De Moivre* et *Laplace* est de démonstration très difficile. Il est admis ici.

Remarque : il ne faut pas confondre l'écart-type $\frac{\sigma}{\sqrt{n}}$ de la variable aléatoire $\overline{X}$ (qui est définie sur l'ensemble des échantillons possibles de taille n) avec l'écart-type d'un échantillon prélevé. L'écart-type de l'échantillon prélevé n'interviendra pas dans nos calculs dans cette partie. Pour éviter cette confusion, la quantité $\frac{\sigma}{\sqrt{n}}$ sera parfois appelée "*erreur type*".

Exemple :

Les statistiques des notes obtenues en mathématiques au BAC STI en France pour l'année 2006 sont :

Moyenne nationale : μ =10,44

Écart-type : σ = 1,46

Une classe de BTS comporte 35 élèves en 2006/2007 issus d'un BAC STI en 2006.

Calculer la probabilité que la moyenne de cette classe soit supérieure à 10.

Ici, nous ne connaissons pas la loi sur la population, mais l'effectif n de l'échantillon est supérieur à 30.

Nous allons donc pouvoir utiliser le T.C.L. 2.

Notons $\overline{X}$ la variable aléatoire qui, à tout échantillon de taille $n = 35$, fait correspondre sa moyenne.

Alors : $$\overline{X} \rightsquigarrow N\left(\mu ; \frac{\sigma}{\sqrt{n}}\right) = N\left(10,44 ; \frac{1,46}{\sqrt{35}}\right)$$

Posons $T = \dfrac{\overline{X} - 10,44}{\dfrac{1,46}{\sqrt{35}}}$ ainsi $T \rightsquigarrow N(0 ; 1)$.

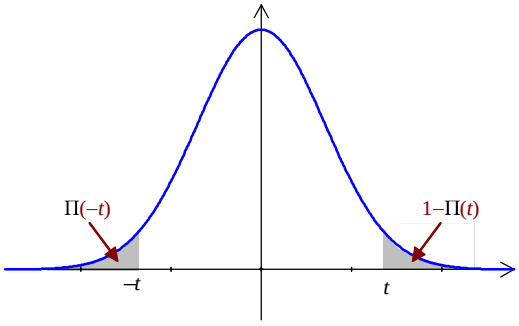
Nous obtenons alors par centrage et réduction :

$$P(\overline{X} \geqslant 10) = P\left(\frac{\overline{X} - 10,44}{\dfrac{1,46}{\sqrt{35}}} \geqslant \frac{10 - 10,44}{\dfrac{1,46}{\sqrt{35}}}\right)$$

$$= P(T \geqslant -1,78)$$

$$= P(T \leqslant 1,78)$$

$$= \Pi(1,78)$$



Remarque : $P(T \geqslant t) = P(T \leqslant -t)$

En effet :

$P(T \geqslant t) = 1 - P(T \leqslant t) = 1 - \Pi(t) = \Pi(-t) = P(T \leqslant -t)$

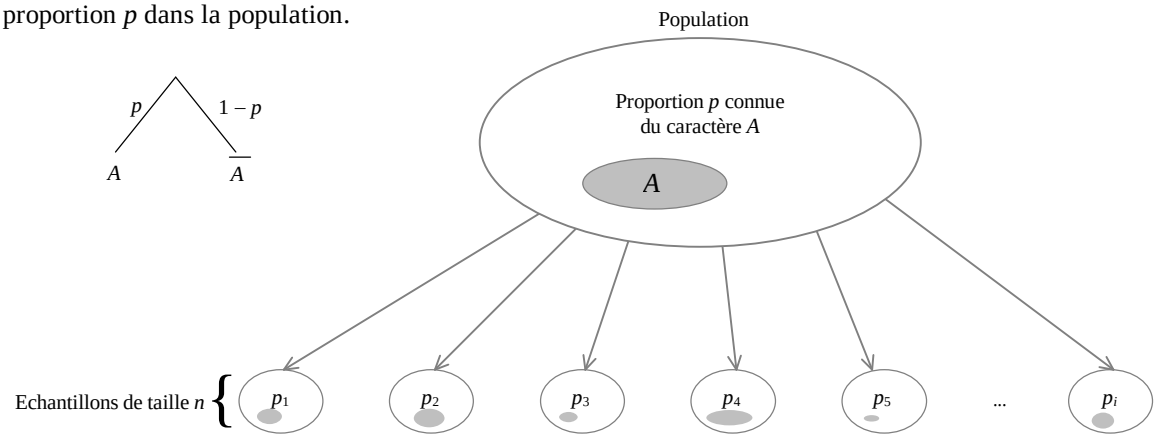
Et par lecture directe de la table de la loi normale centrée-réduite :

$$\Pi(1,78) = 0,9625$$

Conclusion : il y a environ 96% de chance que, dans cette classe de BTS, la moyenne des notes au baccalauréat de Mathématiques soit supérieure à 10.

A.2. Étude d'une proportion dans un échantillon

Cette fois-ci, on dispose d'une population sur laquelle on étudie un caractère (ou attribut) A dont on connaît la proportion p dans la population.



On s'intéresse aux échantillons de taille n. La proportion du caractère A dans les échantillons sera-t-elle toujours la même ? Evidemment non, cette proportion varie en fonction de l'échantillon choisi. Notons F la variable aléatoire qui, à chaque échantillon de taille n, associe sa proportion du caractère A (F s'appelle *distribution des fréquence des échantillons*). Que peut-on dire de cette variable aléatoire F ?

Théorème

Contexte : une population sur laquelle on étudie un caractère A répandu avec une fréquence p.

On prélève, au hasard, un échantillon (tirages avec remise ou assimilés) de taille n **avec** $n \geqslant 30$.

On note F la fréquence du caractère A dans l'échantillon.

Alors la variable aléatoire F suit **approximativement** une loi normale :

$$F \rightsquigarrow N\left(p ; \sqrt{\frac{p(1-p)}{n}}\right)$$

Démonstration :

Nous allons avoir ici un modèle binomial ou apparenté dont on sait qu'il converge vers la loi normale.

Pour tout i compris entre 1 et n, notons X_i la variable aléatoire définie par :

$$X_i = \begin{cases} 1 \text{ si le } i\text{-ème élément de l'échantillon possède l'attribut } A \\ 0 \text{ sinon} \end{cases}$$

La variable aléatoire X_i suit une loi de Bernoulli de paramètre p.

La variable aléatoire $X = X_1 + X_2 + ... + X_n$ est donc binomiale de paramètres n et p :

$$X \rightsquigarrow B(n, p)$$

En conséquence : $$E(X) = np \text{ et } \sigma(X) = \sqrt{np(1-p)}$$

La variable aléatoire $F = \frac{X}{n}$ correpond ainsi à la fréquence de l'attribut A dans l'échantillon.

D'après les propriétés de l'espérance et de l'écart-type :

$$E(F) = \frac{E(X)}{n} = p \text{ et } \sigma(F) = \frac{\sigma(X)}{n} = \sqrt{\frac{p(1-p)}{n}}$$

Exemple :

Une élection **a eu lieu** et un candidat a eu 40 % des voix.

On prélève un échantillon de 100 bulletins de vote.

Quelle est la probabilité que, dans l'échantillon, le candidat ait entre 35 % et 45 % des voix ?

Ici, nous avons $n = 100$ et $p = 0,4$. La variable aléatoire F correspondant à la fréquence des votes pour le candidat dans l'échantillon vérifie donc :

$$F \rightsquigarrow N\left(0,4 ; \sqrt{\frac{0,4 \times 0,6}{100}}\right) = N\left(0,4 ; \frac{\sqrt{0,24}}{10}\right)$$

Posons $T = \dfrac{F - 0,4}{\dfrac{\sqrt{0,24}}{10}}$ ainsi $T \rightsquigarrow N(0 ; 1)$. Nous obtenons alors par centrage et réduction :

$$P(0,35 \leqslant F \leqslant 0,45) = P(-1,02 \leqslant T \leqslant 1,02) = 2\Pi(1,02) - 1$$

Et par lecture directe de la table de la loi normale centrée-réduite $\Pi(1,02) = 0,8461$.

D'où : $$P(0,35 \leqslant F \leqslant 0,45) = 0,6922$$

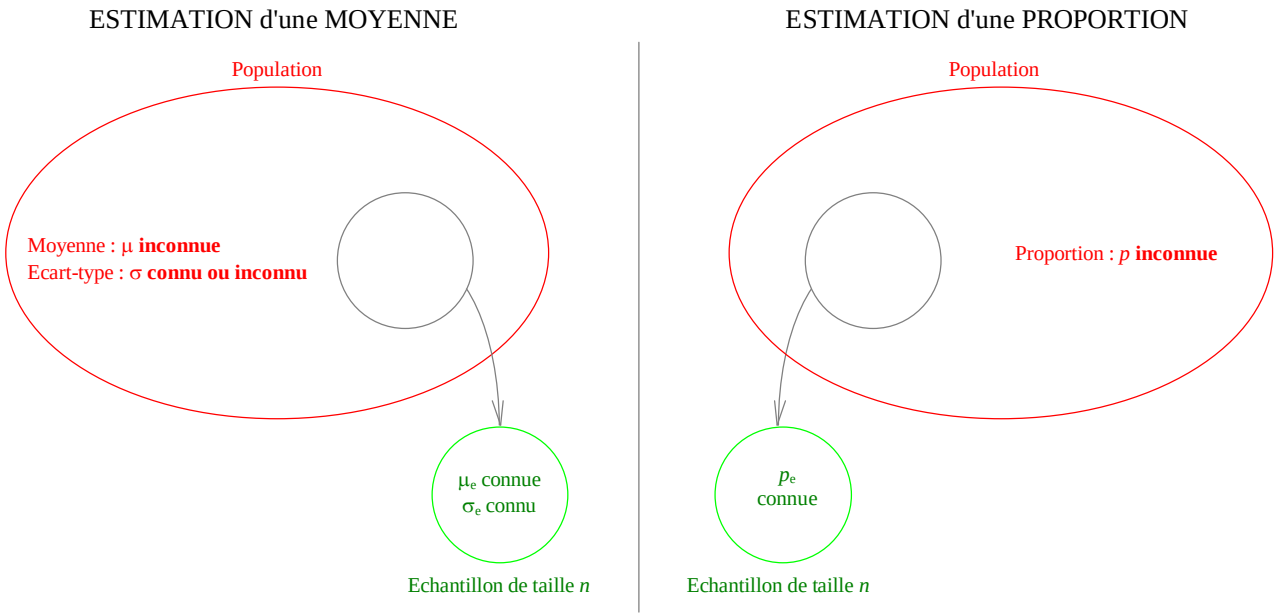
Il y a donc environ 69 % de chance que, dans un échantillon de taille $n = 100$, le candidat ait entre 35 % et 45 % des voix.

En analysant l'exercice ci-dessus, on constate que l'on dispose des informations sur la population (ici, l'ensemble des votes) parce que l'élection a déjà eu lieu. On en déduit des informations sur l'échantillon. Mais, dans la pratique, c'est souvent le phénomène réciproque que nous étudierons : les élections n'ont pas encore eu lieu et on voudrait retrouver les informations sur la population grâce un sondage réalisé sur un échantillon. D'où la deuxième partie de ce document consacrée à l'estimation.

- Partie B - Estimation -

L'objectif de cette partie est de répondre à la problématique suivante : *comment, à partir d'informations (couple moyenne/écart-type ou proportion) calculées sur un échantillon, retrouver ou plutôt estimer celles d'une population entière ?* L'estimation est le problème réciproque de l'échantillonnage. (Mais nous aurons besoin des résultats établis sur la théorie de l'échantillonnage pour passer à la phase estimative).

Nous distinguerons deux cas : celui où l'on cherche à estimer la moyenne μ d'une variable aléatoire définie sur une population et celui où l'on cherche à estimer la proportion d'individus p ayant tel caractère dans la population.



B.1. Estimation d'une moyenne

B.1.1. Estimation ponctuelle

Contexte : on considère une variable aléatoire X sur une population de moyenne (ou espérance) μ inconnue et d'écart-type σ inconnu (ou connu). On suppose que l'on a prélevé un échantillon de taille n (tirage avec remise ou assimilé) sur lequel on a calculé la moyenne μ_e et l'écart-type σ_e.

Une estimation ponctuelle $\widehat{\mu}$ de la moyenne μ de la population est :

$$\widehat{\mu} = \mu_e$$

Une estimation ponctuelle $\widehat{\sigma}$ de l'écart-type σ_e de la population est :

$$\widehat{\sigma} = \sqrt{\frac{n}{n-1}}\,\sigma_e$$

Le coefficient $\sqrt{\frac{n}{n-1}}$ s'appelle *correction de biais*. Lorsque la taille n de l'échantillon est assez grand (en pratique $n \geqslant 30$), ce coefficient est très voisin de 1, si bien que, dans ce cas, on peut estimer $\hat{\sigma} \simeq \sigma_e$.

Exemple :

Une université comporte 1500 étudiants. On mesure la taille de 20 d'entre eux. La moyenne μ_e et l'écart-type σ_e calculés à partir de cet échantillon sont :

$$\mu_e = 176 \text{ cm} \quad \text{et} \quad \sigma_e = 6 \text{ cm}$$

Nous pouvons donc estimer les paramètres de la population :

$$\hat{\mu} = 176 \text{ cm} \quad \text{et} \quad \hat{\sigma} = \sqrt{\frac{20}{19}} \times 6 \simeq 6{,}16 \text{ cm}$$

Remarque :

Nous n'avons fait qu'une estimation, il est bien sûr impossible de retrouver les vraies caractéristiques μ et σ de la population.

L'estimation ponctuelle permet surtout de disposer d'une valeur de référence pour poursuivre/affiner les calculs. On souhaiterait notamment pouvoir faire une estimation par intervalle, en contrôlant le risque pris.

B.1.2. Estimation par intervalle de confiance

Le contexte est le même que le précédent, sauf que nous allons raisonner en deux temps, une phase *a priori* (ou prévisionnelle) dans lequelle on suppose que l'échantillon n'est pas encore prélevé et une phase *a posteriori* dans laquelle on suppose connue la moyenne μ_e et l'écart-type σ_e de l'échantillon et donc la moyenne estimée $\hat{\mu}$ et l'écart-type estimé $\hat{\sigma}$ de la population.

- PHASE *A PRIORI* - Mise en place du modèle prévisionnel -

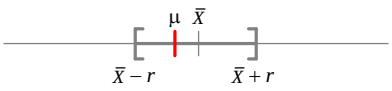
Nous avons vu, dans la théorie sur l'échantillonnage, que si $\overline{X}$ est la variable aléatoire correspondant à la moyenne d'un échantillon de taille n pris au hasard, alors le Théorème Central Limite permet d'affirmer que $\overline{X}$ suit approximativement une loi normale :

$$\overline{X} \rightsquigarrow N\left(\mu ; \frac{\sigma}{\sqrt{n}}\right)$$

Nous allons chercher un intervalle qui contient μ avec une confiance arbitraire de 95% (cela pourrait aussi être 99% ou un autre coefficient de confiance). Nous cherchons donc un rayon r tel que :

Probabilité que la moyenne μ de la population tombe dans un intervalle du type $[\overline{X} - r ; \overline{X} + r]$

$$P(\overline{X} - r \leqslant \mu \leqslant \overline{X} + r) = 0{,}95$$



Cette disposition des inégalités n'est pas pratique mais il y a une correspondance remarquable entre deux événements qui va nous faciliter les calculs :

$$\bar{X} - r \leqslant \mu \leqslant \bar{X} + r$$

Retranchons $\bar{X}$ et μ dans chaque membre :

$$-\mu - r \leqslant -\bar{X} \leqslant r - \mu$$

Multiplions par –1 : $r + \mu \geqslant \bar{X} \geqslant \mu - r$

Remettons les inégalités dans l'ordre croissant :

$$\mu - r \leqslant \bar{X} \leqslant r + \mu$$

Nous sommes ainsi ramenés à calculer :

Cette propriété découle de la symétrie de la valeur absolue :

$$|X - Y| \leqslant r$$

Cela signifie que l'écart entre X et Y est inférieur à r, ce qui s'écrit indifféremment :

$$-r \leqslant X - Y \leqslant r$$

$$Y - r \leqslant X \leqslant Y + r$$

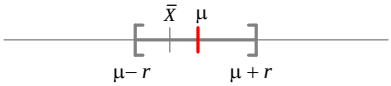
Ou encore :

$$-r \leqslant Y - X \leqslant r$$

$$X - r \leqslant Y \leqslant X + r$$

Probabilité que la moyenne $\bar{X}$ de l'échantillon tombe dans un intervalle centré en μ.

$$P(\mu - r \leqslant \bar{X} \leqslant \mu + r) = 0{,}95$$



Dans la pratique, nous partirons de cette écriture pour déterminer un intervalle de confiance.

On sait que la variable aléatoire $T = \dfrac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{\sqrt{n}}{\sigma}\left(\bar{X} - \mu\right)$ suit la loi normale centrée-réduite $N(0 ; 1)$.

Nous obtenons donc, par centrage et réduction :

$$P\left(\frac{\mu - r - \mu}{\frac{\sigma}{\sqrt{n}}} \leqslant \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leqslant \frac{\mu + r - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = 0{,}95$$

On constate ici que le fait de ne pas connaître μ n'est pas gênant, à ce stade.

$$P\left(-\frac{r\sqrt{n}}{\sigma} \leqslant T \leqslant \frac{r\sqrt{n}}{\sigma}\right) = 0{,}95$$

$$P\left(-\frac{r\sqrt{n}}{\sigma} \leqslant T \leqslant \frac{r\sqrt{n}}{\sigma}\right) = 0{,}95$$

$$2\Pi\left(\frac{r\sqrt{n}}{\sigma}\right) - 1 = 0{,}95$$

$$\Pi\left(\frac{r\sqrt{n}}{\sigma}\right) = 0{,}975$$

Rappel : si $T \rightsquigarrow N(0 ; 1)$ alors :

$$P(-\alpha \leqslant T \leqslant \alpha) = 2\Pi(\alpha) - 1$$

En effet :

$$P(-\alpha \leqslant T \leqslant \alpha) = \Pi(\alpha) - \Pi(-\alpha)$$

$$= \Pi(\alpha) - (1 - \Pi(\alpha))$$

$$= 2\Pi(\alpha) - 1$$

$$\Pi(t) = 0{,}975 \text{ où } t = \frac{r\sqrt{n}}{\sigma}$$

Nous cherchons donc, par lecture inverse de la table de la loi normale centrée réduite une borne t telle que :

$$\Pi(t) = 0{,}975$$

La borne $t = 1{,}96$ convient.

La borne t dépend du coefficient de confiance choisi.

Avec un coefficient de confiance de 99%, nous aurions obtenu :

$$2\Pi\left(\frac{r\sqrt{n}}{\sigma}\right) - 1 = 0{,}99$$

$$\Pi(t) = 0{,}995$$

$$t = 2{,}575$$

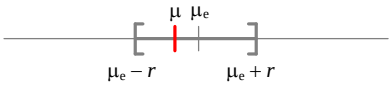
Par la suite, nous noterons t le réel tel que $2\Pi(t) - 1 = C$ où C est le degré de confiance choisi.

Ainsi, notre réel r recherché est tel que : $\dfrac{r\sqrt{n}}{\sigma} = t$

Le rayon r de l'intervalle cherché est : $\boxed{r = t\dfrac{\sigma}{\sqrt{n}}}$

- PHASE *A POSTERIORI* - Utilisation des valeurs estimées ponctuellement -

Nous supposons maintenant que l'échantillon a été tiré, nous obtenons donc **une représentation μ_e de la variable aléatoire** $\overline{X}$:



Nous pouvons affirmer que l'intervalle obtenu **pour cet échantillon**

$$\left[\mu_e - t\frac{\sigma}{\sqrt{n}} \ ; \ \mu_e + t\frac{\sigma}{\sqrt{n}}\right]$$

fait partie d'une famille dans laquelle 95 % contiennent la vraie moyenne μ de la population. On l'appelle **intervalle de confiance à 95 %** (ou autre selon le coefficient de confiance décidé préalablement).

Pour calculer les bornes de cet intervalle, deux cas de figure se présentent selon que nous connaissons ou pas l'écart-type σ de la population. S'il est connu, il n'y a rien à faire :

$$IC = \left[\mu_e - t\frac{\sigma}{\sqrt{n}} \ ; \ \mu_e + t\frac{\sigma}{\sqrt{n}}\right]$$

Si l'écart-type σ de la population n'est pas connu, on le remplace par son estimation ponctuelle $\widehat{\sigma} = \sqrt{\dfrac{n}{n-1}}\,\sigma_e$.

Dans ce cas, nous obtenons : $r = t\sqrt{\dfrac{n}{n-1}}\dfrac{\sigma_e}{\sqrt{n}} = t\dfrac{\sigma_e}{\sqrt{n-1}}$

Nous pouvons donc estimer avec une confiance de 95 % (ou 99 % selon le cas) que la moyenne μ de la population appartient à l'intervalle :

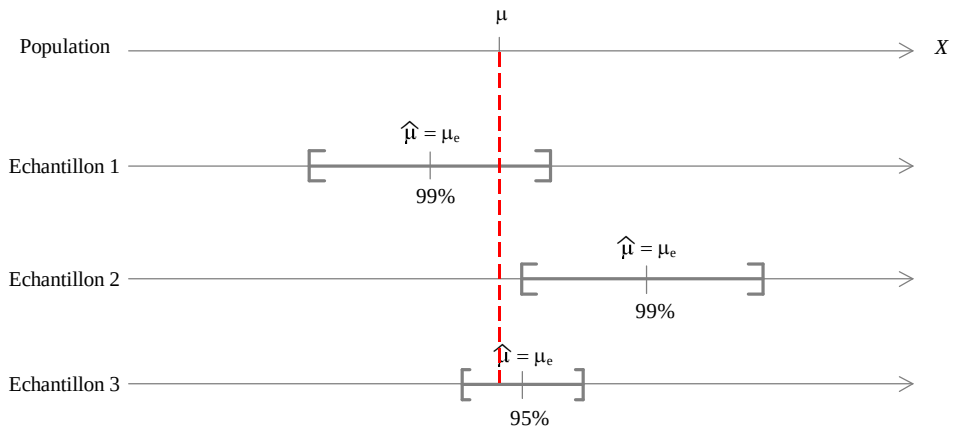
$$IC = \left[\mu_e - t\frac{\sigma_e}{\sqrt{n-1}} \ ; \ \mu_e + t\frac{\sigma_e}{\sqrt{n-1}}\right]$$

On ne retiendra pas cette formule.
Dans la pratique, on refait les calculs.

Remarques :

- L'intervalle de confiance est centré en la valeur μ_e car c'est la seule valeur de référence que nous disposons.
- Le centre de l'intervalle de confiance (à savoir μ_e) dépend de l'échantillon choisi (puisque μ_e en dépend). Son rayon en dépend aussi lorsqu'on ne connaît pas l'écart-type de la population.
- La vraie valeur μ de la moyenne de la population peut ne pas appartenir à l'intervalle de confiance.
- Le rayon de l'intervalle de confiance (à savoir la quantité $r = t\dfrac{\sigma}{\sqrt{n}}$) dépend du degré de confiance C choisi. Plus le degré de confiance C est proche de 100%, et plus la borne t sera élevée et donc le rayon grand.

Illustration :



Un intervalle de confiance ne contient pas forcément la moyenne μ de la population.

Un intervalle de confiance à 95 % est plus petit qu'un intervalle de confiance à 99%. Il risque moins de contenir la valeur moyenne μ.

Exemple :

Une université comporte 1500 étudiants. On mesure la taille de 20 d'entre eux. La moyenne μ_e et l'écart-type σ_e calculés à partir de cet échantillon sont :

$$\mu_e = 176 \text{ cm} \quad \text{et} \quad \sigma_e = 6 \text{ cm}$$

Nous avons déjà estimé ponctuellement les paramètres de la population :

$$\hat{\mu} = 176 \text{ cm} \quad \text{et} \quad \hat{\sigma} = \sqrt{\frac{20}{19}} \times 6 \simeq 6{,}16 \text{ cm}$$

Déterminons maintenant une estimation de μ par intervalle de confiance à 95% (ou au risque de 5 %).

Notons $\overline{X}$ la variable aléatoire correspondant à la moyenne d'un échantillon de taille 20 pris au hasard.

Nous savons que : $\overline{X} \rightsquigarrow N\left(\mu\,;\,\frac{\sigma}{\sqrt{n}}\right) = N\left(\mu\,;\,\frac{\sigma}{\sqrt{20}}\right)$

On calcule un rayon r tel que : $P(\mu - r \leqslant \overline{X} \leqslant \mu + r) = 0{,}95$

On pose $T = \dfrac{\overline{X} - r}{\frac{\sigma}{\sqrt{20}}}$, ainsi T suit la loi normale centrée-réduite $N(0\,;\,1)$.

Nous avons donc : $P\left(-\frac{r\sqrt{20}}{\sigma} \leqslant T \leqslant \frac{r\sqrt{20}}{\sigma}\right) = 0{,}95$

$$2\Pi\left(\frac{r\sqrt{20}}{\sigma}\right) - 1 = 0{,}95$$

$$\Pi\left(\frac{r\sqrt{20}}{\sigma}\right) = 0{,}975$$

$$\Pi(t) = 0{,}975 \quad \text{où } t = \frac{r\sqrt{20}}{\sigma}$$

Nous cherchons donc, par lecture inverse de la table de la loi normale centrée réduite une borne t telle que :

$$\Pi(t) = 0{,}975$$

La borne $t = 1{,}96$ convient.

Ainsi, notre réel r recherché est tel que : $\dfrac{r\sqrt{20}}{\sigma} = 1{,}96$

$$r = \frac{1,96 \times \sigma}{\sqrt{20}}$$

Mais une fois l'échantillon tiré, nous avons obtenu un écart-type estimé $\widehat{\sigma} \simeq 6,16$ cm.

D'où : $r \simeq 2,7$

La réalisation de l'intervalle de confiance à 95% sur cet échantillon est :

$$IC = [176 - 2,7 \ ; \ 176 + 2,7]$$

$$IC = [173,3 \ ; \ 178,7]$$

Nous pouvons donc estimer, avec une confiance de 95 % que la taille moyenne de la population est comprise entre 173,3 cm et 178,7 cm.

B.2. Estimation d'une proportion

B.2.1. Estimation ponctuelle

Contexte : on considère un caractère (ou attribut) A sur une population dont la proportion p est inconnue. On suppose que l'on a prélevé un échantillon de taille n (tirage avec remise ou assimilé) sur lequel on a calculé la proportion p_e d'individus ayant le caractère A.

Notons F la variable aléatoire correspondant à la proportion du caractère A dans un échantillon de taille n pris au hasard. On rappelle qu'alors F suit approximativement une loi normale :

$$F \rightsquigarrow N\left(p \ ; \sigma_p\right) \text{ où } \sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

Une estimation ponctuelle $\widehat{p}$ de la proportion p de l'attribut A dans la population est :

$$\widehat{p} = p_e$$

Une estimation ponctuelle $\widehat{\sigma_p}$ de l'écart-type σ_p est selon le cas :

$$\sqrt{\frac{n}{n-1}} \sqrt{\frac{p_e(1-p_e)}{n}} = \sqrt{\frac{p_e(1-p_e)}{n-1}} \text{ si } n \leqslant 30$$

Correction de biais.

$$\sqrt{\frac{p_e(1-p_e)}{n}} \text{ si } n > 30$$

$$\sqrt{\frac{1}{4n}} \text{ si statisticien pessimiste}$$

Ces estimations ponctuelles de l'écart-type ne sont pas utiles dans l'immédiat. Elle serviront pour la détermination d'un intervalle de confiance de la proportion.

Exemple :

À quelques jours d'une élection, un candidat fait effectuer un sondage. Sur les 150 personnes interrogées, 45 se disent prêtes à voter pour lui aux prochaines élections.

La proportion d'individus prête à voter pour ce candidat dans l'échantillon est ici de $p_e = \frac{45}{150} = 0,3$.

On estime donc qu'il en est de même dans la population (comment pourrait-on faire autrement ?) :

$$\widehat{p} = p_e = 0,3$$

Quand à l'indication σ_p, on peut ici l'estimer par :

$$\widehat{\sigma_p} = \sqrt{\frac{p_e(1-p_e)}{n}} = \sqrt{\frac{0,3 \times 0,7}{150}} \simeq 0,037$$

On voudrait aller plus loin et, au lieu d'une simple proportion, calculer un intervalle contenant, avec une confiance arbitraire fixée au départ, la proportion p d'individus prêts à voter pour ce candidat.

B.2.2. Estimation par intervalle de confiance

Le contexte est le même que le précédent. Nous avons vu, dans la théorie sur l'échantillonnage, que si F est la variable aléatoire correspondant à la proportion d'un caractère dans un échantillon de taille n pris au hasard, alors F suit approximativement une loi normale :

$$F \rightsquigarrow N\left(p\,;\sigma_p\right) \text{ où } \sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

Nous avons déjà remarqué que le fait que p soit inconnu n'est pas gênant dans les calculs *a priori*. Le problème ici, c'est que nous ne connaissons pas l'écart-type $\sqrt{\frac{p(1-p)}{n}}$. Nous le remplaçerons, dans la phase *a posteriori*, par son estimation ponctuelle (qui est $\sqrt{\frac{p_e(1-p_e)}{n-1}}$ en général ou $\sqrt{\frac{p_e(1-p_e)}{n}}$ si la correction de biais n'est pas proposée ou encore $\sqrt{\frac{1}{4n}}$ si nous voulons une hypothèse pessimiste).

Cherchons un intervalle qui contient p avec une confiance arbitraire de 90 % (cela pourrait être un autre coefficient de confiance). Nous cherchons donc un rayon r tel que :

$$P(F - r \leqslant p \leqslant F + r) = 0{,}90$$

Nous avons déjà vu que cette probabilité pouvait s'écrire de manière plus pratique :

$$\boxed{P(p - r \leqslant F \leqslant p + r) = 0{,}90}$$

On sait que la variable aléatoire $T = \frac{F - p}{\sigma_p}$ suit la loi normale centrée réduite $N(0\,;\,1)$.

Nous obtenons donc, par centrage et réduction :

$$P\left(\frac{p - r - p}{\sigma_p} \leqslant \frac{F - p}{\sigma_p} \leqslant \frac{p + r - p}{\sigma_p}\right) = 0{,}90$$

$$P\left(\frac{-r}{\sigma_p} \leqslant T \leqslant \frac{r}{\sigma_p}\right) = 0{,}90$$

$$2\Pi\left(\frac{r}{\sigma_p}\right) - 1 = 0{,}90$$

$$\Pi\left(\frac{r}{\sigma_p}\right) = 0{,}95$$

On cherche une borne t telle que : $\Pi(t) = 0{,}95$ avec $t = \frac{r}{\sigma_p}$

Par lecture inverse de la table de la loi normale centrée réduite $N(0\,;\,1)$:

$$t = 1{,}645$$

Ce qui nous permet de calculer r : $r = t\,\sigma_p$

Supposons maintenant l'échantillon prélevé. Nous avons donc une estimation pontuelle de p et σ_p.

Ainsi, la réalisation de l'intervalle de confiance dans l'échantillon est :

$$IC = \left[p_e - t\sqrt{\frac{p_e(1-p_e)}{n-1}} \ ; \ p_e + t\sqrt{\frac{p_e(1-p_e)}{n-1}} \right]$$

On ne retiendra pas cette formule. Dans la pratique, on refait les calculs.

Remarques :

- Si on n'effectue pas la correction de biais, l'intervalle de confiance est :

$$IC = \left[p_e - t\sqrt{\frac{p_e(1-p_e)}{n}} \ ; \ p_e + t\sqrt{\frac{p_e(1-p_e)}{n}} \right]$$

- On peut également se placer dans une hypothèse pessimiste en choisissant un écart-type maximal. Nous savons que la parabole d'équation $y = x(1 - x)$ admet un maximum égal à $\frac{1}{4}$ en $\frac{1}{2}$.

  Ainsi l'écart-type maximal est $\sqrt{\frac{1}{4n}}$. Il a, de plus, l'avantage d'être indépendant de p.

  Dans ce cas, la réalisation de l'intervalle de confiance dans l'échantillon est :

$$IC = \left[p_e - t\sqrt{\frac{1}{4n}} \ ; \ p_e + t\sqrt{\frac{1}{4n}} \right]$$

Exemple :

A quelques jours d'une élection, un candidat fait faire un sondage. Sur les 150 personnes interrogées, 45 se disent prêtes à voter pour lui aux prochaines élections.

La proportion d'individus prête à voter pour ce candidat dans l'échantillon est ici de $p_e = \frac{45}{150} = 0{,}3$.

On a déjà estimé ponctuellement : $\widehat{p} = p_e = 0{,}3$ et $\widehat{\sigma_p} \simeq 0{,}037$

Déterminons maintenant une estimation de p par intervalle de confiance à 80%.

Notons F la variable aléatoire correspondant à la proportion d'individus prêts à voter pour ce candidat dans un échantillon de taille 150 pris au hasard.

Nous avons vu qu'approximativement :

$$F \rightsquigarrow N(p\,;\sigma_p) \quad \text{où } \sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

On cherche un rayon r tel que : $P(p - r \leqslant F \leqslant p + r) = 0{,}8$

$$2\Pi\left(\frac{r}{\sigma_p}\right) - 1 = 0{,}8$$

$$\Pi\left(\frac{r}{\sigma_p}\right) = 0{,}9$$

Par lecture inverse de la table de la loi normale centrée-réduite, on cherche une borne t telle que :

$$\Pi(t) = 0{,}9 \ \text{ avec } t = \frac{r}{\sigma_p}$$

La valeur $t \simeq 1{,}28$ convient donc : $r = 1{,}28\ \sigma_p$

Supposons maintenant l'échantillon prélevé. Une estimation ponctuelle de σ_p est $\widehat{\sigma_p} \simeq 0{,}037$.

D'où : $r \simeq 0{,}047$

La réalisation de l'intervalle de confiance dans cet échantillon est alors

$$IC = [0{,}3 - 0{,}047 \ ; \ 0{,}3 + 0{,}047]$$

$$IC = [0{,}253 \ ; \ 0{,}347]$$

$$IC_{\%} = [25{,}3 \ ; \ 34{,}7]$$

Nous pouvons estimer, avec une confiance de 80 %, que la proportion d'individus dans la population prêts à voter pour le candidat en question est comprise entre 25,3 % et 34,7 %.

Exercice :

Une usine fabrique des câbles. Un câble est considéré comme conforme si sa résistance à la rupture X est supérieure à 3 tonnes. L'ingénieur responsable de la production voudrait connaître, en moyenne, la résistance à la rupture des câbles fabriqués.

Il n'est, bien sûr, pas question de faire le test sur toute la production (l'usine perdrait toute sa production !).

Un technicien prélève donc un échantillon de 100 câbles dans la production. Notons $\overline{X}$ la variable aléatoire correspondant à la force à exercer sur le câble pour le rompre. Le technicien obtient les résultats suivants :

$$E(\overline{X}) = 3{,}5 \text{ tonnes}$$

$$\sigma(\overline{X}) = 0{,}4 \text{ tonne}$$

Proportion de câbles dont la résistance est supérieure à 3 tonnes : $p_e = 0{,}85$

1. a. Donner une estimation ponctuelle de la moyenne μ et de l'écart-type σ de la variable aléatoire X dans la production.
   b. Déterminer une estimation par intervalle de confiance à 95 % de la moyenne μ de X.
2. a. Donner une estimation ponctuelle de la proportion p de câbles conformes dans la production.
   b. Déterminer une estimation par intervalle de confiance à 90 % de cette proportion.

- RÉSUMÉ -

- Echantillonnage -

MOYENNE

Echantillons de taille n de moyenne $\bar{X}$

$$\bar{X} \rightsquigarrow N\left(\mu\ ;\ \frac{\sigma}{\sqrt{n}}\right)$$

où μ et σ sont la moyenne et l'écart-type dans la population.

PROPORTION

Echantillons de taille n avec une fréquence F

$$F \rightsquigarrow N\left(p ; \sqrt{\frac{p(1-p)}{n}}\right)$$

où p est la proportion dans la population.

- Estimation -

MOYENNE

Population de moyenne μ ***inconnue*** *et d'écart-type σ.*

Echantillon de taille n ***connu*** *de moyenne μ_e et d'écart-type σ_e.*

☞ Une estimation ponctuelle de μ est μ_e.

☞ Une estimation ponctuelle de σ est $\sqrt{\frac{n}{n-1}}\,\sigma_e$ si $n \leqslant 30$ ou tout simplement σ_e sinon ($n > 30$).

☞ Pour estimer μ par intervalle avec une confiance C (par ex 95%), on cherche un rayon r tel que :

$$P(\mu - r \leqslant \bar{X} \leqslant \mu + r) = C$$

où $\bar{X} \rightsquigarrow N\left(\mu\ ;\ \frac{\sigma}{\sqrt{n}}\right)$

On exprime r en fonction de σ et on remplace σ par sa valeur connue ou son estimation ponctuelle.

$$IC = [\mu_e - r\ ;\ \mu_e + r]$$

PROPORTION

Proportion ***inconnue p*** *dans une population.*

Echantillon de taille n ***connu*** *avec une proportion p_e.*

☞ Une estimation ponctuelle de p est p_e.

☞ Une estimation ponctuelle de σ_p est $\sqrt{\frac{p_e(1-p_e)}{n-1}}$ si $n \leqslant 30$ ou $\sqrt{\frac{p_e(1-p_e)}{n}}$ sinon ($n > 30$).

☞ Pour estimer p par intervalle avec une confiance C (par ex 95%), on cherche un rayon r tel que :

$$P(p - r \leqslant F \leqslant p + r) = C$$

où $F \rightsquigarrow N(p\ ;\sigma_p)$ avec $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$

On exprime r en fonction de σ_p et on remplace σ_p par son estimation ponctuelle.

$$IC = [p_e - r\ ;\ p_e + r]$$

Coefficient de confiance	80 %	90 %	95 %	99 %
Valeur de $\Pi(t)$	0,9	0,95	0,975	0,995
Borne t	1,28	1,645	1,96	2,575